CE's Sample Design

David Swanson
Statistical Methods Division
Bureau of Labor Statistics

Overview

- Selection of PSUs
- Sample size (national and PSU-level)
- Selection of addresses within PSUs
- 2006 sample cut
- Quality measures
- Improvements

New Sample of Geographic Areas and Addresses Selected Every Decade

- 2010 Census-Based Sample Design (in development)
- 2000 Census-Based Sample Design (2005–2014)
- 1990 Census-Based Sample Design (1996–2004)
- 1980 Census-Based Sample Design (1986–1995)

BLS

Randomness Ensures Representativeness

- CE is a nationwide household survey.
- Geographic areas randomly selected to represent the total U.S.
- Households randomly selected to represent the geographic areas.
- Systematic sampling of households ensures every segment of the population is well-represented.

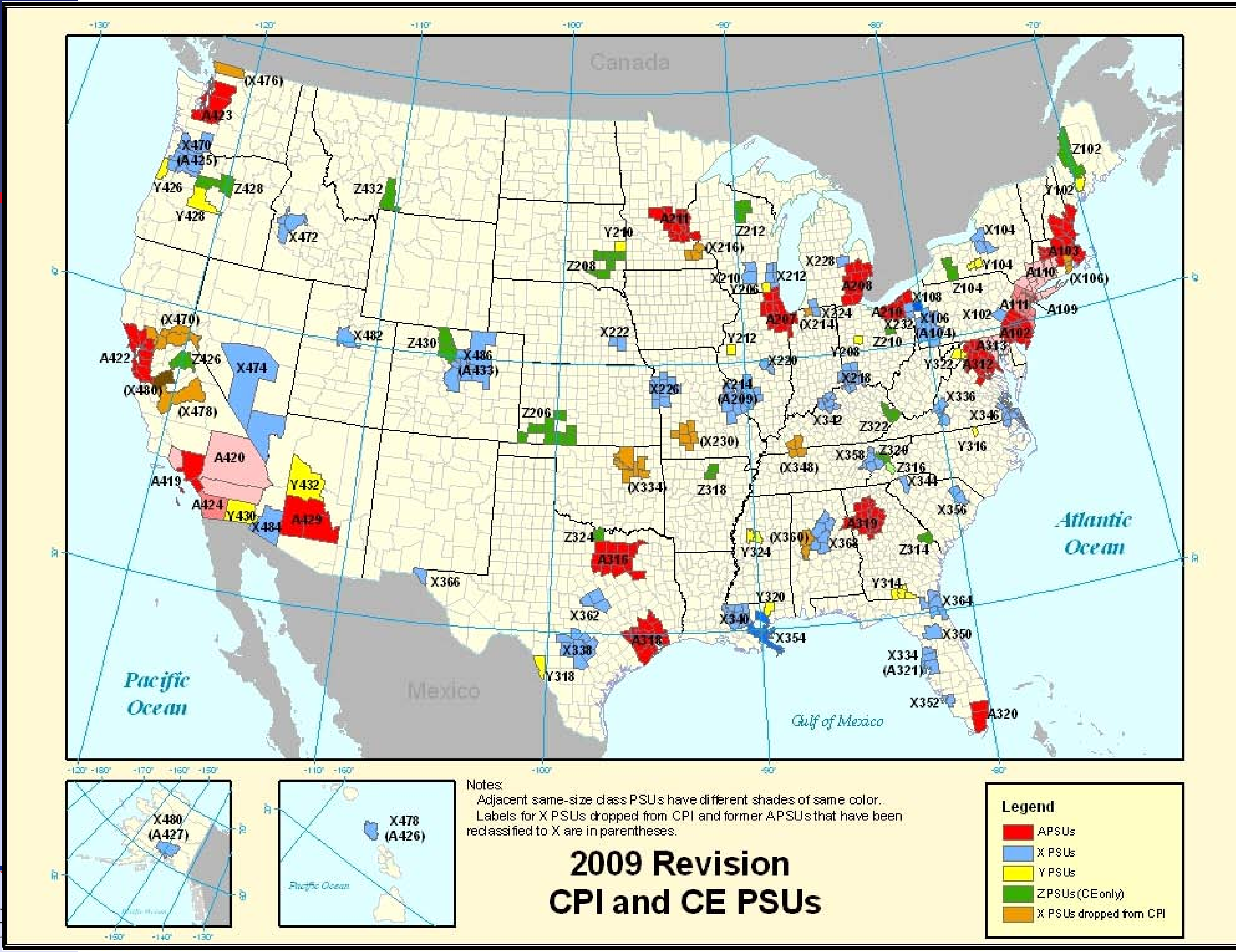
Geographic Sample (2000 Sample Design)

- 28 "A" PSUs – metropolitan CBSAs over 2 million people
- 42 "X" PSUs – metropolitan CBSAs under 2 million people
- 16 "Y" PSUs – micropolitan CBSAs
- 16 "Z" PSUs – non-CBSA ("rural") areas

"A" PSUs are self-representing

"X," "Y," "Z" PSUs are non-self-representing

BLS



Canada
Mexico
Pacific Ocean
Atlantic Ocean
Gulf of Mexico
(X476)
A423
X470
(A425)
Y426
Z428
Y428
Z432
X472
A211
Y210
Z212
(X216)
Z208
Z102
Y102
X104
A103
Y104
A110
(X106)
Z104
X228
X210
X212
Y206
A208
A207
(X214)
X224
A210
X108
X232
X106
(A104)
X102
A111
A109
A102
Z210
Y208
A313
Y322
A312
(X470)
A422
Z426
X474
(X480)
(X478)
X482
Z430
X486
(A433)
X222
Y212
X220
X226
X214
(A209)
X218
X336
X342
Z322
X346
Z206
(X230)
A420
A419
Y432
A424
Y430
X484
A429
(X334)
Z318
(X348)
X358
Z320
Z316
X344
Y316
X356
A319
Z324
A316
Y324
(X360)
X362
Z314
X366
Y314
Y320
X364
X362
X340
X354
A318
X338
X350
X334
(A321)
Y318
X352
A320
X480
(A427)
X478
(A426)
Notes:
Adjacent same-size class PSUs have different shades of same color.
Labels for X PSUs dropped from CPI and former APSUs that have been reclassified to X are in parentheses.
2009 Revision
CPI and CE PSUs
Legend
APSUs
X PSUs
Y PSUs
Z PSUs (CE only)
X PSUs dropped from CPI

Non-Self-Representing PSUs

X, Y, Z PSUs are grouped into "strata" according to a 5-variable geographic model:

1. Latitude
2. Longitude
3. Latitude squared
4. Longitude squared
5. Percent urban

One PSU is randomly selected from each stratum (pps)

BLS

PSUs in Stratum X344

PSU	Population
Charlotte, NC-SC	1,114,808
Charleston-North Charleston, SC	549,033
✓**Greenville, SC**	**379,616**
Fayetteville-Fort Bragg, NC	302,963
Columbus, GA-AL	274,624
Gastonia, NC	190,365
Wheeling, WV-OH	153,172
Warner Robbins, GA	134,433
Total	**3,099,014**

BLS

Household Sample - National

Target Sample Size

- 7,700 interviewed households per year (Diary)
- 7,700 interviewed households per quarter (Interview, interviews 2-5 only)

Target Sample Yield

- 15,400 weekly diaries per year (=7,700×2)
- 30,800 quarterly interviews per year (=7,700×4)

Household Sample - Local

Local Target Sample Size

- Allocate 7,700 interviewed households to individual PSUs proportional to each stratum's population (pps)
  - Sample allocated to 40 CPI_AREAs (28 "A" PSUs + 12 region-size classes), then sub-allocated to individual PSUs
  - Minimum of 80 interviewed households per CPI_AREA
  - Urban areas over-sampled (and rural areas under-sampled) to help CPI
- Minimizes CE's nationwide variance

BLS

Translate Interviewed Households into Addresses

- 80% "eligibility" rate
- 75% response rate
- 60% "participation" rate (0.60 = 0.80×0.75)

BLS

Translate Interviewed Households into Addresses (continued)

PSU	Interviewed households	Addresses
A102 Philadelphia	169	322
A103 Boston	195	286
A104 Pittsburgh	80	123
A109 New York City	220	420
A110 NY-Conn suburbs	212	335
A111 New Jersey suburbs	182	291
etc.	*etc.*	*etc.*
Total	**7,700**	**12,800**

BLS

Select a Random Sample of Addresses

Census Bureau's "100% Detail File"

Sort from poor to rich (information from 2000 Census)

- Urban/rural
- Number of people in household
- Tenure (owner, renter)
- Market value of home (owners)
- Monthly rent (renters)

BLS

Select a Random Sample of Addresses (continued)

Compute the sampling interval for each PSU

Sampling interval = (# addresses in sampling frame) ÷ (# addresses in CE sample)

Typical sampling intervals:

- every 5,000th address (A PSUs)
- every 1,000th address (X,Y,Z PSUs)

Select a Systematic Sample of Addresses (continued)

-- D --- I --- D --- I --- D --- I --- D -- I --- D --- I --- D --- I --- *etc.*

D=Diary, I=Interview

- 12 years of addresses selected
- Reserve sample

Sampling Frames

- Unit frame (85%)
- Area frame (10%)
- Permit frame (5%)
- Group Quarters (<1%)

2006 Sample Cut

7 PSUs changed from "A" to "X"

11 "X" PSUs cut from sample

8% of addresses cut from sample

New Geographic Sample (2006-present)

- 21 "A" PSUs – metropolitan CBSAs over 2.7 million people
- 38 "X" PSUs – metropolitan CBSAs under 2.7 million people
- 16 "Y" PSUs – micropolitan CBSAs
- 16 "Z" PSUs – non-CBSA ("rural") areas

BLS

New Sample Size (2006 – present)

Target Sample Size

- 7,050 interviewed households per year (Diary)
- 7,050 interviewed households per quarter (Interview, interviews 2-5 only)

Target Sample Yield

- 14,100 weekly diaries per year (=7,050×2)
- 28,200 quarterly interviews per year (=7,050×4)

BLS

Sample Design Improvements

Recent

- Cluster sampling research
- PSU stratification
- "Optimal" allocation of nationwide sample to individual PSUs
- Nonresponse bias

Future

- Annual Sampling
- Master Address File (MAF)

BLS

Estimation

BLS

Overview

- Basic estimator
- Weights
- Data adjustments (allocation, imputation)
- Accuracy of estimates (Variances, PCE)

Average Annual Expenditure per Consumer Unit

Item	2007	2008	2009
All Items	$49,638	$50,486	$49,067
Food	6,133	6,443	6,372
Housing	16,920	17,109	16,895
Apparel and Services	1,881	1,801	1,725
Transportation	8,758	8,604	7,658
etc.			

BLS

CE's Basic Estimator is a Weighted Average

$$\bar{x}_i = \frac{\sum_c w_c x_{ic}}{\sum_c w_c}$$

where

w_c = weight of consumer unit "c"

x_{ic} = expenditure of CU=c for item=i

BLS

Weights

Final weight

= (Base weight) $\times$ (Nonresponse adjustment factor) $\times$ (Calibration adjustment factor)

Typical values:

$15{,}000 = 10{,}000 \times 1.33 \times 1.13$

Weights: Nonresponse Adjustment

Nonresponse adjustment factor

= (# occupied housing units) ÷ (# responders)

64 demographic groups (64=4×4×2×2)

- 4 regions (Northeast, Midwest, South, West)
- 4 CU sizes (1,2,3-4,5+)
- 2 tenures (owner, renter)
- 2 races (Black, Non-Black)

Typical factor = 1.33

BLS

Weights: Calibration Adjustment

Calibration adjustment factor

= (CPS population estimate) ÷ (CE population estimate)

CE population estimate

= (# responders) × (# people per CU) × (nonresponse adjustment factor)

24 population controls from CPS (24=14+8+2)

- 7 Age × 2 Race categories
- 4 Regions × 2 Urban/Rural
- 2 Tenure (owner/renter)

Typical factor = 1.13

BLS

Data Adjustments

Allocation – Split combined expenditures into individual item categories

Example:

Reported: $100 clothing

Allocated: $60 shoes, $20 skirts, $20 pants

Imputation – Fill-in unreported values

Allocation

1. Specific items mentioned, but not specific expenditures (e.g., $100 for shoes, skirts, pants):

Allocate the expenditure to the items reported based on the nationwide distribution of expenditures (e.g., 60% shoes, 20% skirts, 20% pants)

Allocation (continued)

2. Specific items not mentioned (e.g., $100 for clothing)

2a. Identify all item categories whose median expenditure is below the reported expenditure

2b. Randomly select one of those item categories and assign the median expenditure to the item category

2c. Subtract the assigned amount from the reported expenditure

2c. Repeat 2b and 2c until the reported expenditure is allocated completely

BLS

Imputation Methods

- Percentage distributions
- Weighting class
- Traditional regression estimator

  $(y = \beta_0 + \beta_1 x)$
- Regression with multiple imputation

  $(y_i = \beta_0 + \beta_1 x + \varepsilon_i \quad \text{for } i=1,2,3,4,5)$

BLS

Imputation – Demographic Characteristics

Percentage distribution -- match on 2 or 3 characteristics (e.g., region and income class), then randomly select a value.

IF RANUNI(0) < 0.6750 THEN TENURE='OWNER';
ELSE TENURE='RENTER';

Tenure	Percent	Cumulative percent
Owner	0.6750	0.6750
Renter	0.3250	1.0000

BLS

Imputation - Expenditures

Weighting class – match on 2 or 3 variables (e.g., region and income class), then use the mean expenditure from all reported expenditures in the cell.

Income quartile	Mean reported expenditure for pre-paid phone cards
1	$78.51
2	67.80
3	61.51
4	53.53

BLS

Rental Equivalence

"If someone were to rent this home today, how much do you think it would rent for monthly, unfurnished and without utilities?"

20% of CPI's weight.

Imputation – Rental Equivalence

Old: Use the reported value from the most similar housing unit (hotdeck)

New: Traditional regression estimator

($y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$)

$y = \beta_0 + \beta_1 \cdot \text{Bedrooms} + \beta_2 \cdot \text{Bathrooms} + \beta_3 \cdot \text{Property\_Value} + \ldots$

BLS

Imputation - Income

- Old: None (analysis limited to "complete reporters")
- New: Multiple imputation (5 estimates)
- Cold-deck (5 years of data)
- Backward model selection
- Variable selection & parameter estimation done quarterly

$$\text{Salary} = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Education} + \ldots + \varepsilon_i$$

$$\text{Pension} = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Retired} + \ldots + \varepsilon_i$$

for i=1,2,3,4,5

BLS

Accuracy of Estimates: Standard Errors
CE Standard Errors for 2009

Item	Annual Expenditure	Standard Error	Coefficient of Variation
All Items	$49,067	$595	1.21%
Housing	16,895	194	1.15
Transportation	7,658	165	2.16
Food	6,372	84	1.31
etc.			

SEs are estimated with Balanced Repeated Replication (BRR)

BLS

Quality Measures: Coverage Rate

	2004	2005	2006	2007	2008	2009
Male	88%	87	88	87	86	86
Female	87	86	87	86	86	85
Non-Black	89	89	89	89	88	88
Black	71	72	72	69	72	68

Coverage Rate = (CE population estimate) ÷ (CPS population estimate) × 100%

CE population estimate = (# responders) × (# people per CU) × (nonresponse adjustment factor)

BLS

Estimation Improvements

Recent

- CAPI (fewer outliers & missing values)
- User-Friendly Diary
- Rental Equivalence imputation
- Income imputation

Future

- Coverage of the Black population
- Income tax imputation

BLS